Diatom Identification: a Double Challenge Called ADIAC

Hans du Buf, University of Algarve, Faro, Portugal
Micha Bayer and Stephen Droop, Royal Botanic Garden Edinburgh, U.K.
Ritchie Head and Steve Juggins, University of Newcastle, U.K.
Stefan Fischer and Horst Bunke, University of Berne, Switzerland
Michael Wilkinson and Jos Roerdink, University of Groningen, The Netherlands
José Pech-Pacheco and Gabriel Cristóbal, Instituto de Optica (CSIC), Madrid, Spain
Hamid Shahbazkia and Adrian Ciobanu, University of Algarve, Faro, Portugal

Abstract: *This paper introduces the project ADIAC (Automatic Diatom Identification and Classification), which started in May 1998 and which is financed by the European MAST (Marine Science and Technology) programme. The main goal is to develop algorithms for an automatic identification of diatoms using image information: both valve shape (contour) and ornamentation. The paper presents the goals of the project as well as first results on shape modeling and contour extraction. Public data are available in order to create student projects beyond the ADIAC partnership. For further information see http://www.ualg.pt/adiac*

1 Introduction

ADIAC is the acronym of the project Automatic Diatom Identification and Classification. In phycological research the word identification refers to what in pattern recognition is meant by classification, whereas the phycological meaning of classification is the establishment of the class-forming rules. In order to avoid a confusion, we will apply the phycological interpretation. Since this is the first scientific publication by most of the project's partners, and because our goal is to promote diatom identification as a new and challenging area in pattern recognition, we start with explaining the history of the project, diatom research and the goals. Sections 2 to 4 present first results on image databases, shape modeling and contour extraction.

1.1 History: from hobby to profession

The ADIAC project was "born" in May 1998 but its "conception" took place a few years before at the Musée National de l'Histoire Naturelle de Paris, in Simone Servant's office to be precise, although she seemed not quite aware of the very fact. So what happened? Microscopy and a general interest in Nature's richness in morphology in various areas such as zoology, geology and botany being one of my hobbies, I (HdB) came to collect antique microscope slides, the production of which had become an art that bloomed in the 19th century throughout Europe. Occasionally strolling around Paris in search for slide collections at the few scientific instruments antique shops, I came into contact with Bernard Coupel of the Vaast bookshop, rue Jussieu, who happened to purchase part of the Tempère collection including microscopes and very special ornamental slides made by J. Tempère, one of France's greatest diatomists. Bernard Coupel showed me a few of these artistic preparations with many different diatoms arranged like flowers and beautifully coloured because of the light diffraction, slides that were not for sale of course given their uniqueness [1]. He advised me to see Simone Servant at the nearby museum. She told me about the diatom collections there, the research and the existence of a journal named Diatom Research. I had seen a computer in her office and because of my background a very natural question that I asked her was about the application of computers to diatom recognition. I cannot remember her precise answer but I grasped that she did not really know what I was talking about, but then and there I had an idea that I could not put aside; an idea that has now materialised into a three-year, seven-partner and 1.25 million Euro project with a main funding from the DG XII's MAST programme.

1.2 What are diatoms?

Diatoms are unicellular algae related to brown algae (*Phaeophyta*, e.g. seaweeds like *Fucus* and *Laminaria*), yellow-green (*Xanthophyta*) and golden-brown (*Chrysophyta*) algae, but not at all related to red, green or blue-green algae. Almost all need sunlight to grow, and live almost anywhere where there is enough light and moisture: in the water column of the sea, lakes and rivers, in sediments underneath and at the edge of wa-

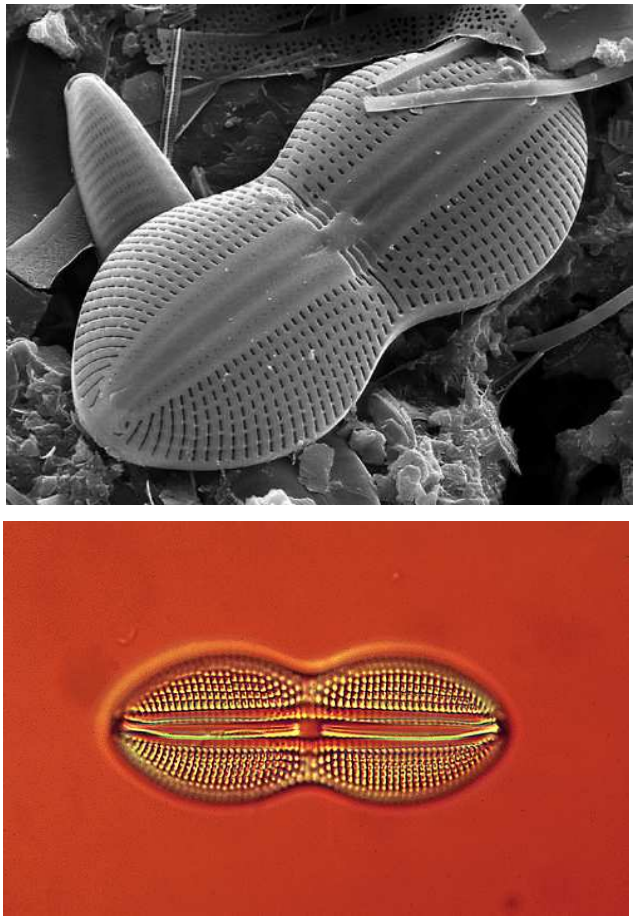

Figure 1: SEM (scanning electron microscope) image of *Diploneis heemskerkiana* (top) and the optical microscopy (DIC) equivalent (bottom).

ter bodies, and also near the surface of damp soils. Estimates vary widely, but there may be as many as 200,000 species in the world [3], making them the second most diverse group of plants after the flowering ones. They are ecologically very important and contribute around 20% of the world's carbon fixation, which makes them more productive than all the world's rainforests.

There are three aspects of diatom biology that make them important well beyond their intrinsic appeal (see Fig. 1) and their contribution to world ecology. First, they have a cell wall made of silica that is very resistant to decay – the cell walls can survive in lake- and seabed sediments for thousands and even millions of years after the cell itself has died. Second, the ornamentation of the cell wall is highly specific, and most diatoms can be identified to species level on the basis of the cell wall alone. Third, each diatom species tends to be able to survive and grow only in a relatively narrow range of ecological conditions. These three factors together mean that diatoms can be used for ecological monitoring, for reconstructing past environments, and in archaeological, geological and forensic research.

1.3 Diatom research

Diatom-based research ranges from purely systematic (classification and evolutionary studies) to purely applied, where diatoms are used as an analytical tool in ecological, geological, climatological, geographical, archaeological or forensic research. There are no clear boundaries, but many of those who use diatoms as a tool have more interest in (and are more skilled in) the application for which they are using the diatoms than in the diatoms themselves. In addition, many of these applications require identification of a large range of diatom species. This combination of factors (identification of many species by non-specialists) has led to a need for quick and effective identification aids. Unfortunately, written floras are not adequate, since they are slow to produce and update as new information becomes available and as classifications change. A computer-based system, however, would have many advantages: identification would no longer require the level of human expertise that it does at present (although it would still require human expertise to interpret the computers' results), and a computer-based system could be kept completely up-to-date with respect to the publication of new species and other changes in the classification, especially if the reference database were kept centrally and accessed via the Web.

1.4 ADIAC goals

The main goal is of course the development of a complete software system for a completely unsupervised diatom identification using only image information. This aim is going too far because there are thousands of different taxa and the identification rules are sometimes not very clear. Also, in some cases one valve view is not enough and additional information is required. Nevertheless, ADIAC is the first European project devoted to diatom identification on the basis of both valve contour and ornamentation. Explicit goals are: (1) to develop image databases with different discrimination complexities, (2) to develop methods for an automatic slide scanning on microscopes with motorised stages, (3) to develop methods for obtaining a complete, graphical, diatom-valve description, (4) to develop an identification system using for example graph matching that can produce a sorted list with best matches, (5) to test all methods using the image databases and (6) to integrate the methods into taxonomic and ecological database systems.

The second goal is illustrated by Fig. 2, which shows a low-magnification overview of a strewn slide with many diatoms and biological debris. Only the automi-

Figure 2: Low-magnification image of a strewn slide.

sation of the scanning process, i.e. by marking only the positions of almost complete diatoms and a subsequent high-magnification image capture at each marked position using also autofocusing, would already result in a tremendous saving of labour. The saving of labour by the complete processing as proposed and studied by ADIAC would be much more, and most researchers involved with diatoms could spend much more time doing their own work in geology, climatology, etc. Therefore, the ultimate goal will be to install analysis and identification software together with a huge diatom database on one or more central servers, which would allow researchers to send images by email and to receive automatically a sorted list of best matches. In other words, ADIAC and subsequent efforts are expected to carry diatom research well into the 20th century.

1.5 An open pilot study

During the three ADIAC years it is expected to realise state-of-the-art algorithms and huge image databases. However, it is the first project in which diatom contour and ornamentation will be explored for an identification. Hence, it is expected that the project will need a continuation to further improve the analysis tools, but mainly to establish image databases that contain all diatom species relevant for certain applications/sites. It is therefore extremely important that ADIAC creates a database with potential users by contacting researchers who could profit from the project. Because this is a very time-consuming task, we also invite individual researchers, institutions and companies to contact us and to let us know what their applications are and what they would expect. ADIAC will organise several workshops to which interested researchers will be invited, and we hope to establish active collaborations beyond the ADIAC project partnership. Images, publications, references etc can be found in the ADIAC webpages at URL http://www.ualg.pt/adiac and mirrored at http://www.rbge.org.uk/adiac. These contain all addresses, telephone numbers and related webpages of all partners plus additional links. Contacts, for example for establishing student projects or participation in ADIAC workshops and meetings, are *not* limited to the Coordinator. Main contact: Hans du Buf (coordinator), University of Algarve, Vision Laboratory, Faculty of Exact Sciences and Humanities, Campus de Gambelas - UCEH, 8000 Faro, Portugal; Tel: +351 89 800900 ext 7761; Fax: +351 89 818560; Email: dubuf@ualg.pt

2 Image databases

A database of ca 1200 digital images has already been created, which is likely to grow to ca 10,000 by the end of the project. Images are produced using a number of techniques: most are captured directly from the microscope by one of two digital cameras; others are photographed using monochrome film and the developed negatives are scanned using a slide-scanner.

Irrespective of their size, diatoms are captured using the maximum magnification that will allow for the entire specimen to be photographed at a resolution of 10 pixels per micron or better. This resolution is more or less the minimum that can capture all the resolution of which an optical microscope is capable.

Photographs of diatoms usually include extraneous material or illumination artefacts that detract from their quality. The use of digital media means that such imperfections can be removed relatively easily (depending on their severity), providing their source is understood. Preprocessing involves removal of imperfections from the two main sources: those that are part of the magnification, illumination and imaging systems (and nothing to do with the specimens themselves), and those that are intrinsic to the specimen preparations (such as other diatoms, girdle bands or other debris that interfere with the diatom to be photographed).

Ultimately, the storage of images will be in conjunction with a taxonomic database which is being developed for specialised diatom use [2]. Until the structure of the database is completed, images are kept in more or less unstructured folders, and the accompanying information (identification, provenance, slide number, microscope stage coordinates, pixel size and shape) is kept in a separate index file. Please refer to the project web site for the full methodology for specimen preparation and image capture.

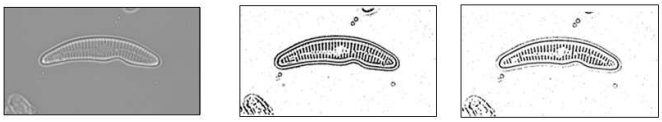

Figure 3: Dark and bright contours in the grey-level image (left), the double contour after a local thresholding (middle) and the double contour eliminated (right).

3 Contour extraction and shape analysis

In a brightfield-microscopic image, a diatom's silica frustule leaves a dark outline. The organic material, that still remains on the frustule after cleaning it, leaves a very light signature (halo) outside the dark contour because of diffraction, see Fig. 3 (left). We use this information in order to develop a low-level and data-oriented contour extraction. The idea is to threshold the grey-level image, to label the connected elements and to follow the external outline to obtain the contour. As the illumination variance is not regular in these images, a local thresholding must be used. However, a drawback of a local thresholding is that the bright part creates a double contour in the binarised image (Fig. 3 middle). To overcome this, first we find the Otsu threshold between the histogram's maximum and the histogram's brightest element. Next we set the value of all pixels above this threshold to the value of the threshold, and only then the local thresholding is applied to the image (Fig. 3 right). After the binarisation the connected elements are labeled and only the most central and largest element is chosen as a first contour candidate.

The contour is obtained by following the external part of the chosen element in the clockwise sense, using a "always turn left" algorithm. If the contour is not closed then the algorithm tries to connect the first element to another element in the neighbourhood. The result goes through a correction process that eliminates concave but thin deformations and that fills in the internal gaps. If necessary the final contour is rotated to a horizontal position by using symmetry axes and moments before analysing the precise shape.

Valve shapes are roughly divided between centric (circular, triangular and square, the latter two normally cusped with rounded vertices) and pennate (elliptical including specific (a)symmetries). A standard approach is to study the Fourier series of the closed contours for discriminating these forms. This approach has been studied before by modeling pennate *Tabellaria* valves with up to 20 Fourier descriptors [4]. Because we don't know *a priori* how well we can separate and describe different pennate shapes with a fixed or variable number of Fourier descriptors, and because it may be better to apply later a best-fitting

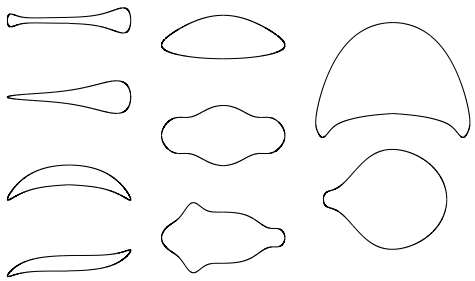

Figure 4: Synthesised diatom contours; see text.

ellipse approach but using other mathematical functions, we studied a few alternatives. As far as we know only fourth-order Cassinian curves and two half ellipses glued together go beyond normal ellipses. Assuming a normalisation in terms of rotation (horizontal shapes) and size, Cassinian curves can approach panduriform shapes with two parameters and glued ellipses can approach semilanceolate shapes also with two parameters.

Coming back to the first pass in the Fourier shape analysis, i.e. the construction of two discrete parametric and periodic signals $X(t)$ and $Y(t)$ following an ellipse in a horizontal position, these signals are pure sines with a phase offset. We noticed that (a)symmetric deformations from a pure ellipse, as found with many pennate diatom shapes, affect only $Y(t)$ and the differences with a pure ellipse can be described by adding symmetrically Gaussian functions or derivatives of Gaussians to $Y(t)$. After some experiments we found that many shapes can be described with a small number of parameters. An example is the sigmoid lanceolate shape, which can be modeled by $X(t) = a\cos(t)$ and $Y(t) = b\sin(t) + c\{\exp(-t^2/d) - \exp(-(t-\pi)^2/d)$. Figure 4 shows some synthesised forms. These are, with the number of parameters between parentheses: left column top-to-bottom: bilobate (8), clavate (5), crescentic (5) and sigmoid lanceolate (5); middle column: semilanceolate (5), without name (8) and without name (6); right column: auricular (10) and spatulate (5). In conclusion, it may be possible to apply a best-fitting ellipse algorithm, followed by a best-fitting *other-shape-with-few-parameters* in order to determine the best parameters to be used in the identification process.

4 Diatom isolation

As can be seen in Fig. 2, in many cases diatoms are connected to debris or partially occluded. Hence, we need algorithms that can isolate individual diatoms and their contours. Contour extraction is done in two steps. In a preprocessing step initial contours are extracted using a conventional edge-following algorithm

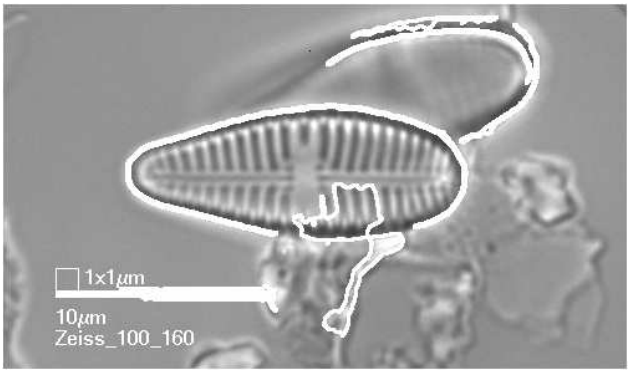

Figure 5: Initial contours of two overlapping diatoms.

like Canny's. Then the object contours are extracted by using the best-fitting ellipse and a subsequent contour following in the elliptical polar-transformed image [5,6].

Figure 5 shows an edge-detection result of an image that contains two diatoms and debris plus a scale bar. Only the initial contours containing more than 200 pixels are shown. Due to small gaps in the edge image the diatom in the centre of the image is described by two contours. The contour of the second diatom is only partly detected since it is not sharply focused and occluded.

Beginning with the contour of maximum length, the best-fitting ellipse is determined for each initial contour. A general conic can be represented by an implicit second order polynomial $G(\mathbf{a},\mathbf{x}) = \mathbf{a}\cdot\mathbf{x} = Ax^2 + By^2 + Cxy + Dx + Ey + F = 0$. $G(\mathbf{a},\mathbf{x}_i)$ is called the algebraic distance of a point (x,y) to the conic $G(\mathbf{a},\mathbf{x}) = 0$. The fitting of a general conic can be done by minimising the sum of squared algebraic distances $\arg\min_{\mathbf{a}} \left\{ \sum_{i=1}^{N} G(\mathbf{a},\mathbf{x}_i)^2 \right\}$ of the curve to the N data points $\mathbf{x}_i$. In order to force the conic to be an ellipse, the constraint $B^2 - 4AC < 0$ has to be fulfilled. The ellipse-fitting problem can be solved by using a generalised eigensystem. The polynomial coefficients $A,\ldots,F$ are computed by determining the eigenvector corresponding to the smallest eigenvalue. For the computation of the polar-transformed image the parametric ellipse form $x = x_c + a\cos\alpha$ and $y = y_c + b\sin\alpha$ is relevant. Here (x_c, y_c) represents the centroid, a, b the radii and α the rotation angle. These parameters can be calculated using the polynomial coefficients. Results of ellipse fitting for the diatoms in the example image are shown in Fig. 6.

Using the parameters of the best-fitting ellipse, the elliptical polar coordinate transform is computed. This polar transform is based on sampling counterclockwise the original gray level image along scan beams with length r. The transform is done in discrete rotation

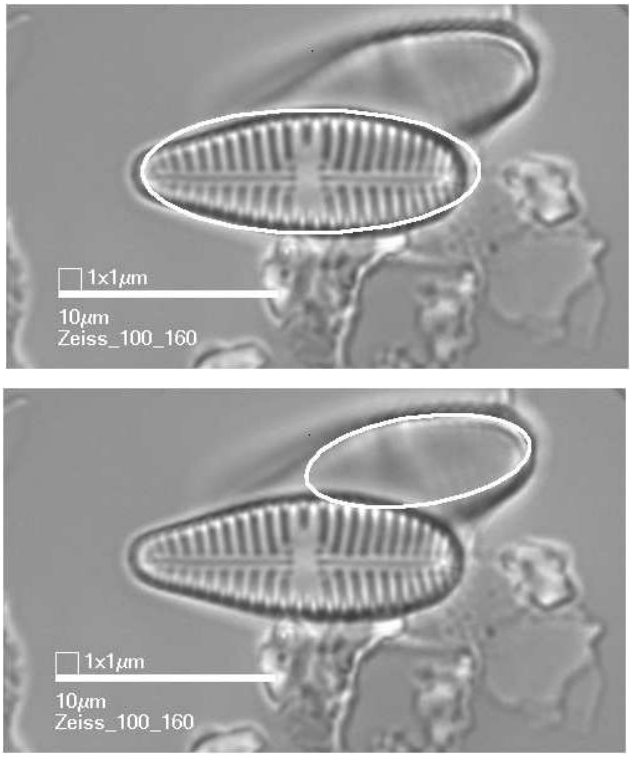

Figure 6: Examples of best-fitting ellipses.

steps $\Delta h = 1^o$. The polar-transformed image coordinates (r,h) are computed using the image coordinates (x,y) according to $x = (m/a)\cdot r\cdot\cos h$ and $y = (m/b)\cdot r\cdot\sin h$, where m is used to normalise the position of the contour. In practice m is fixed to one quarter of the maximum distance from the center of the ellipse to the borders of the image. Furthermore, to compensate for the rotation of the ellipse, each point is rotated around the angle α using $x_r = x\cos\alpha - y\sin\alpha + x_c$ and $y_r = x\sin\alpha + y\cos\alpha + y_c$.

In the polar-transformed image the problem of contour extraction reduces to the extraction of a nearly straight line from the top to the bottom. We apply a depth-first search algorithm which evaluates the gray level changes along the path. One result can be seen in Fig. 7. The back-transformed contours of the isolated diatoms are shown in Fig. 8. Both diatom contours have been accurately isolated.

5 Conclusions

We have presented a brief introduction to diatom research and the ADIAC project, as well as first results on contour extraction and shape analysis. Actually, there is much more progress, but six pages are not sufficient to present all work done; please refer to the webpages. ADIAC being a first pilot project to study a completely computerised diatom identification on the basis of im-

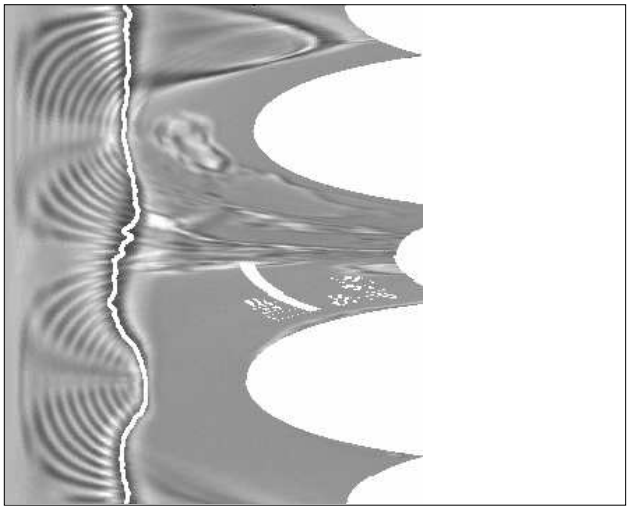

Figure 7: Contour detection in a polar-transformed image around the centre of the best-fitting ellipse.

age information, it is very likely to be continued with the aims of further improving identification accuracy and establishing huge databases on one or more central servers.

Because most information, including image databases and identification rules, even image analysis and identification software, will become publicly available, we hope that many MSc and PhD projects will be created, as has happened before in the area of fingerprint analysis. This will be a fruitful *collaboration* as well as *competition,* i.e. the best algorithms will survive. But an international competition is not a problem because it serves only one goal: the creation of the best identification system which facilitates research in all diatom applications.

Finally, that an automatic diatom identification is a new challenge in pattern recognition will not be a surprise, given the complexity and diversity of the diatom shapes and ornamentations. But why is ADIAC a *double* challenge? As in many pattern-recognition problems, the sometimes very subtle differences in shape and ornamentation between different taxa hamper a clear distinction, even for diatomists. In other words, the *second challenge* is to try to establish exact class-forming rules, which is a non-trivial task in phycology given the inexact way that biological classifications work: i.e. to be able to make an identification even if some of the classificatory information is absent, misobserved or contradictory.

Acknowledgement: ADIAC is funded by the European MAST (Marine Science and Technology) programme, contract MAS3-CT97-0122.

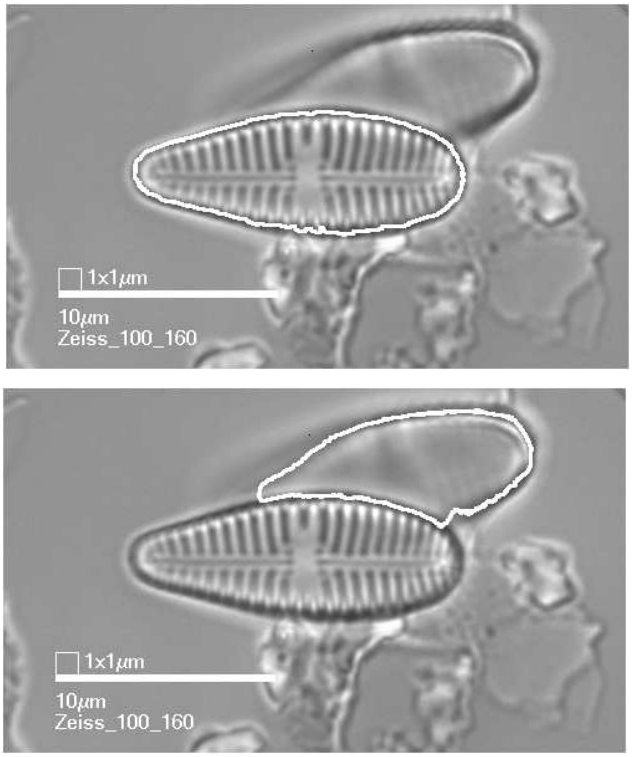

Figure 8: Back-transformed contours of the isolated diatoms.

References

[1] Champreux, F. (1989) Les diatomées et la diatomite. Minéraux et Fossiles, Vol. 15 No. 169, pp. 7-15.

[2] Droop, S.J.M., Sims, P.A., Mann, D.G. and Pankhurst, R.J. (1993) A taxonomic database and linked iconograph for diatoms. In: van Dam, H. (ed.) Proc. Twelfth Int. Diatom Symposium, Renesse, The Netherlands, 30 August - 5 September 1992. Hydrobiologia, Vol. 269/270, pp. 503-508.

[3] Mann, D.G. and Droop, S.J.M. (1996) Biodiversity, biogeography and conservation of diatoms. In: Kristiansen, J. (ed.) Biogeography of Freshwater Algae. Hydrobiologia, Vol. 336, pp. 19-32.

[4] Mou, D. and Stoermer, E.F. (1992) Separating *Tabellaria* (Bacillariophyceae) shape groups based on Fourier descriptors. J. of Phycology, Vol. 28 (3), pp. 386-395.

[5] Niemann, H., Bunke, H., Hofmann, I., Sagerer, G. Wolf, F. and Feistel, H. (1985) A knowledge based system for analysis of gated blood pool studies. PAMI 7(3), pp. 246-259.

[6] Puli, M., Fitzgibbon, A.W. and Fisher, R.B. (1996) Ellipse-specific direct least-squares fitting. IEEE International Conference on Image Processing.